High Performance Computing in Accelerator Physics*

Kwok Ko, Stanford Linear Accelerator Center
kwok@slac.stanford.edu

High performance computing powered by increased investments in software and hardware infrastructures is beginning to have a significant impact on accelerator design and analysis. The US DOE funded Accelerator Grand Challenge has laid the groundwork for developing a suite of electromagnetic codes that are based on unstructured grids and utilize parallel processing on supercomputers such as the Cray T3E and IBM SP2 as well as PC clusters. We will show how this new capability has enabled some of the most challenging problems in accelerator modeling to be solved in resolution, accuracy and turnaround time previously not possible. We will describe the technologies and resources required to support such large-scale simulations, and will discuss the benefits to present and future accelerator facilities from advancing simulation as the third tool of science. Further code development plans under the newly approved DOE SciDAC Initiative will be presented.

INTRODUCTION

The US Department of Energy promotes High Performance Computing (HPC) to help advance the progress of science in its program offices through advanced computing initiatives that include both hardware and software investments. The DOE Grand Challenge was such a program that funded selected teams from various disciplines to tackle the most difficult computational problems in their respective areas. The High Energy and Nuclear Physics (HENP) program office supported the Computational Accelerator Physics team that consisted of two national laboratories (SLAC, LANL) and two universities (Stanford, UCLA). This team focused on two main topics, electromagnetic modeling and beam dynamics simulation, and developed new parallel software to take advantage of massively parallel supercomputers installed at the DOE's National Energy Research Scientific Computing (NERSC) center at Berkeley. The success of the Accelerator Grand Challenge led to the formation of a national collaboration involving six national labs, five universities, and one industrial partner. The funding is provided by DOE's Scientific Discovery through Advanced Computing (SciDAC) initiative whose goal is to foster large multi-institutional, multi-disciplinary teams to develop community codes that run on terascale platforms to help solve the most challenging problems facing the field. The newly approved SciDAC project has been expanded to include a third research area, that of advanced accelerator concepts to study beams under extreme conditions like those found in laser and plasma based accelerators.

Particle accelerators are among the most important and most complex scientific instruments in use, and are critical to research in fields such as high-energy physics, nuclear physics, materials science, chemistry, and the biosciences. They have been proposed for applications that address national needs, and examples include accelerator transmutation of waste, accelerator-driven fission and fusion energy production, accelerator production of tritium, and proton radiography for stockpile stewardship. Smaller scale accelerators have beneficial use in many areas such as irradiation and sterilization of biological hazards, medical isotope production, particle beams for irradiation therapy, ion implantation and beam lithography. Given the great value of particle accelerators it is imperative that the most advanced computing tools and resources be brought to bear on the design and development of these complex facilities and devices. The availability of high performance, large memory parallel supercomputers has made large-scale computing the third tool of scientific discovery complimentary to the traditional approaches of theory and experimentation. Large-scale simulation enables numerical experiments on systems for which physical experimentation would be prohibitively expensive or technologically unfeasible.

This paper will present an overview of the electromagnetic component of the Accelerator Modeling project. Detailed description of codes and results will be covered in several related papers in the session on "High Performance Computing in Accelerator Physics". That session will also include papers addressing the beam dynamics and advanced accelerators areas.

* Work supported by the U.S. DOE Contract No. DE-AC03-76SF00515.

THE NEED FOR HIGH PERFORMANCE COMPUTING

(i) High Resolution Component Design

Accelerator physicists and engineers are faced with increasingly stringent requirements on electromagnetic components as new and existing facilities continually strive towards higher energy and current, and greater efficiency. In the proposed Next Linear Collider (NLC) [1] scheme, the frequency of the accelerating field must be accurate to within 1 part in 10,000 which is comparable to fabrication tolerance in order to maintain acceleration efficicency. This requirement is to be met in a complex cavity geometry that optimizes the accelerating field gradient while suppressing parasitic wafefields generated by the beam. One design, called the Round Damped Detuned Structure (RDDS) is shown in Fig. 1a. Simulating such geometry is challenging for existing electromagnetics software running on desktop computers because it involves a huge number of degrees of freedom (DOF) to model the many curved surfaces to the desired accuracy. While these standard packages have been used extensively by the accelerator community, it became evident that new simulation tools utilizing parallel processing that harness the large memory available in supercomputers were needed to provide the high resolution required for the computer-aided design of these complex accelerating structures. The RDDS cavity as modeled with the parallel eigenmode solver Omega3P consists of one million DOF's is shown in Fig. 1b..

There is another obvious advantage for parallel processing besides gaining access to large memory. In the ideal situation, a parallel code built on scalable algorithms would obtain linear speedup in simulation time. Such a huge jump in processing speed could never be reached by improvement in single CPU performance even if Moore's law continues to prevail. The combination of large memory and scalable computing enables simulation to become a cheaper and faster alternative to the expensive, time-consuming process of repeated fabrication and testing. This has certainly been the case in the design of the RDDS cavity or cell for the NLC. Dimensions generated by simulation were directly used in computer controlled milling machines for fabrication of the cells (Fig. 1c) and subsequent cold–test measurements found frequency accuracy to be close to 0.01 % as predicated by calculation.

(ii) System Scale Analysis

The case for high performance computing in accelerator modeling becomes more compelling when one considers system scale studies of large, heterogeneous structures. For example, the NLC RDDS accelerating section consists of 206 cells each varying from the next by order of microns. This variation in cavity dimensions follows a prescribed distribution so as to detune the dominant higher order modes that constitute the wakefields to reduce their effect. Additional reduction is provided by slot openings in the cavity walls to couple the wakefields out to external manifolds. So far this detuning and damping scheme has only been analyzed using an approximate model that consists of equivalent circuit chains. It is of great interest to model the entire 206-cell section to verify if the desired wakefield suppression can be achieved, and to validate the equivalent circuit model. Such a simulation is estimated to require hundreds of Gigabytes of memory and the calculation of thousands of modes in the long structure. Calculation has been done on structures with fewer cells. Fig. 2a is the Omega3P model of a 47-cell RRDS section and Fig 2b shows for the first time, the actual fields of a localized mode in the structure. Many issues remain to be addressed before full structure simulation can be carried out and they will be described in later sections.

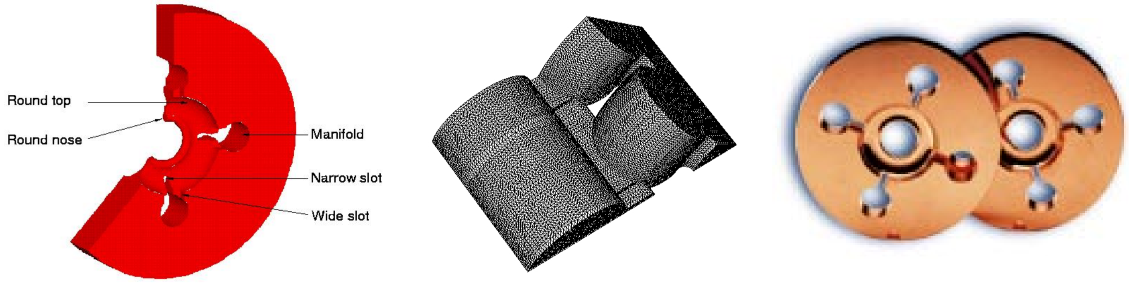

Figure 1a (left) Geometric model of the RDDS cell showing cavity features; Fig. 1b (middle) CUBIT mesh of one and half cell; Fig 1c (right) Fabricated cells using dimensions calculated with Omega3P.

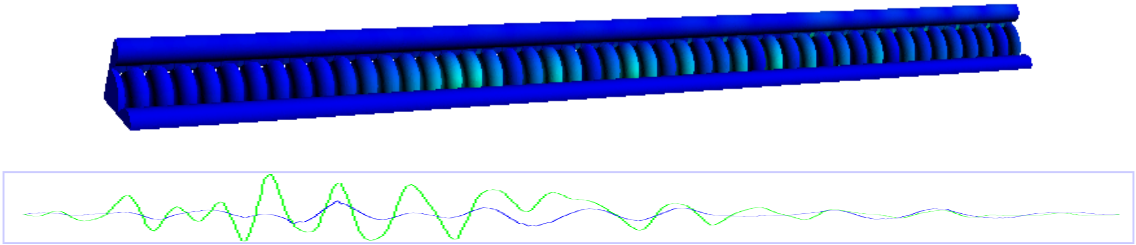

Figure 2a (top) 47-cell section of the NLC RDDS structure simulated with Omega3P; Fig. 2b (bottom) Electric field profiles of a localized mode along the beam axis (green) and in the manifold respectively.

PARALLEL ELECTROMAGNETIC CODE DEVELOPMENT

Motivated by the design needs of the NLC accelerating structure R&D, a parallel electromagnetic code development effort was established in 1997 at SLAC with the award of the DOE Computational Accelerator Physics Grand Challenge. It was determined from the outset that all the codes would be based on 3D unstructured grid to conform to curved surfaces for geometric fidelity. The codes would also share a common geometry input data for mesh distribution and matrix assembly. They would be written in object oriented C++ and their development would follow standardized software practices. Other common features are that they run parallel on distributed memory computers using the Message Passing Interface (MPI), and that they reuse existing parallel libraries as much as possible. For example, the Aztec library is used to perform parallel linear algebra operations, while ParMetis [2] is used to partition the mesh for load balancing. The codes [3] are:

(1) Omega3P – a 3D parallel eigenmode solver to find normal modes in lossless rf cavities using linear and quadratic elements,

(2) Tau3P – a 3D parallel time domain solver to calculate the transmission properties of open structures on a modified Yee grid,

(3) Phi3P – a 3D parallel static solver based on a field formulation that uses hybrid elements for improved accuracy and for more exact description of material boundaries ,

(4) Ptrack – a particle tracking module that works Omega3p and Tau3P to study rf breakdown and dark currents, both important issues for high gradient acceleration.

Omega3P and Tau3P are furthest along in development and are routinely used for modeling complex accelerating structures and beamline components. They have become essential design tools for the NLC linac structure R&D. Omega3P is used to find cavity modes whereas Tau3p is used to match the input and output power couplers to the structure (Fig. 3). Under SciDAC, a complex solver will be developed for Omega3P so that cavities with lossy materials can be modeled while Tau3P will implement a rigid beam for wakefield calculations. Phi3P and Ptrack are new SciDAC projects so therefore are still early in their development. In the following sections the related technologies required for large-scale electromagnetic simulations along with lessons learned as well as proposed research will be described.

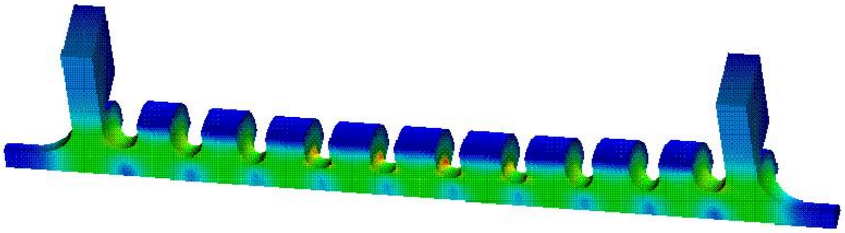

Figure 3. The input coupler for the NLC accelerating structure simulated with Tau3P.

LARGE-SCALE ELECTROMAGNETIC SIMULATIONS

The success of large-scale electromagnetic simulations relies heavily on the combined efforts of a multi-disciplinary team consisting of physicists, applied mathematicians, computer scientists, software engineers, geometry builders and grid generators, and visualization experts. The reason is because much of the computational and software support for parallel computing does not exist or is still being developed. The list includes mesh generation, domain decomposition, solver algorithms, adaptive refinement, and visualization of large datasets. One main goal of the SciDAC initiative is to accelerate the development of the infrastructure for high performance computing to address these issues that are common to other scientific simulations such as computational fluid dynamics and climate modeling. In the case of accelerator modeling, the groundwork has been laid by the Accelerator Grand Challenge project so there is already ongoing research in some of these areas. Through SciDAC, new collaborations are being formed to study those areas that have not previously been addressed. A brief introduction to each topic follows:

(i) Mesh Generation

In order to preserve the realism of the actual structure the mesh generation begins with the construction of a parametric solid model from the engineering drawing as depicted in Figs 4a and 4b. This step usually requires the service of a designer using standard packages such as SolidEdge [4]. The model is able to resolve the small gap spacing between the tuner port and the plunger and to allow the plunger position to vary. An interface is built to connect the solid modeler to the mesh generator. CUBIT [5] is the mesher of choice although other meshers such as SIMAIL have been used as well. The type of meshes to be generated depends on the solver – Omeag3P works with tetrahedral elements whereas Tau3P prefers hexahedral meshes. Fig. 4c shows the Omega3P mesh for the waveguide-damped rf cavity presently operating in the PEP-II [6] storage ring at SLAC. Fig. 5 is the Tau3P mesh for the NLC output coupler to the X-Band accelerating structure. Meshing has always been a lengthy, labor-intensive process for most structural simulations. Two research areas of increased interests are parallel mesh generation and mesh smoothing. The former enables very large meshes to be generated expeditiously while the latter helps solver to improve stability in the time domain (Tau3P) or achieve better convergence in the frequency domain (Omega3P).

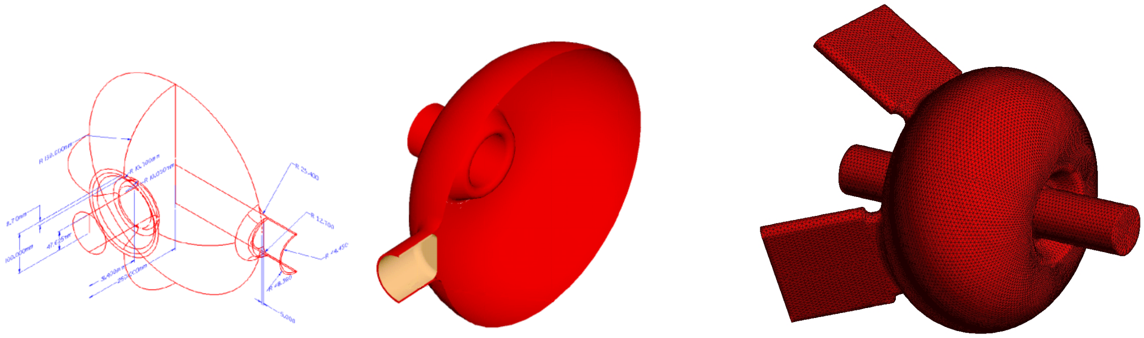

Figure 4a (left) Engineering drawing of PEP-II rf cavity; Fig. 4b (middle) Solid model generated with EMS/Intergraph; Fig. 4c (right) CUBIT mesh with tetrahedral elements.

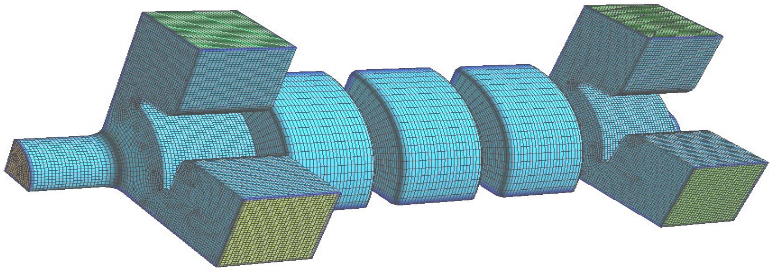

Figure 5. CUBIT mesh with hexahedral elements for the NLC input coupler.

(ii) Domain Decomposition

A key step in parallel computing is domain decomposition or the partitioning of the mesh onto a given number of processors. To handle the mesh distribution a parallel C++ library called DistMesh has been developed that supports a class of elements including tetrahedron, prism, pyramid, and hexahedron. It uses MPI for communication and takes care of the global numbering of mesh quantities such as elements, faces, edges and nodes. DistMesh also replicates interface elements onto the neighboring processors as they are needed by the applications (see Figs. 5b and 5c). Partitioning is improved with the parallel library ParMetis which is an unstructured graph partitioning tool for achieving load balancing (see Fig. 5a). This partitioning procedure has worked well for Omega3P applications as the simulations show low computing cost for communication. It has not worked so well for Tau3P because of the non-orthogonal nature of the dual mesh leading to uneven number of non-zeros across the processors even though the number of elements is balanced. Various schemes are being explored to remedy this situation to improve parallel performance. .

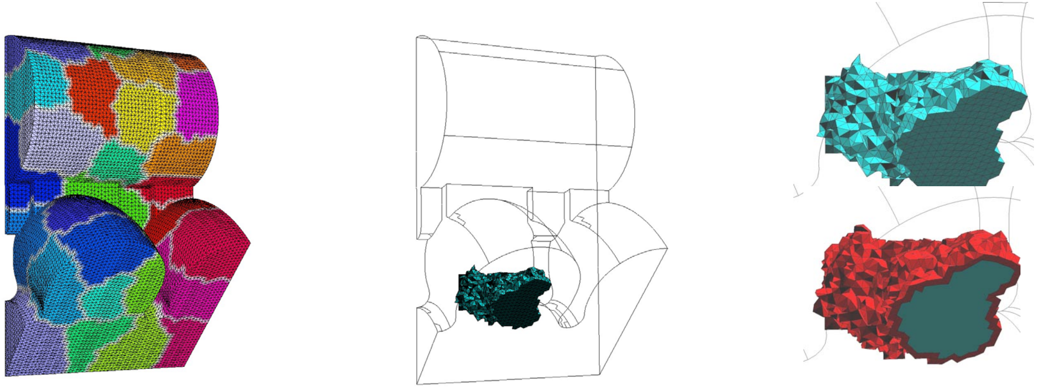

Figure 5a (left) An octant of 1.5 RDDS cell into 32 domains using DistMesh and ParMetis; Fig. 5b (middle) Location of elements on one processor within the mesh; Fig. 5c(right) Local elements on top and replicated elements at bottom for one processor.

(iii) Solver Algorithms

Omega3P solves the generalized eigenvalue problem in parallel for large sparse matrices. The algorithm finds the interior eigenvalues through two steps, a bandpass filtering step by way of inexact shift-invert Lanczos followed by a Jacobi iteration using the Jacobi Orthogonal Component Correction (JOCC) refinement. It makes use of the parallel library Aztec for parallel matrix-vector multiplications. With this solver, Omega3P has been able to treat problem sizes up to tens of millions of DOF's. More importantly, the code is capable of solving cavity frequencies to an accuracy of better than 1 part in 10,000 as required by the NLC RDDS design (Fig. 6a). It also shows good scalability or parallel speedup (Fig. 6b). The performance is only limited by convergence (Fig. 6c) which deteriorates when the eigenvalues are closely clustered. Work under ScidAC is in progress to implement better linear solvers and preconditioners to improve convergence and reduce memory requirement. A more detailed discussion on Omega3P as well as Tau3P can be found in accompanying papers in this conference.

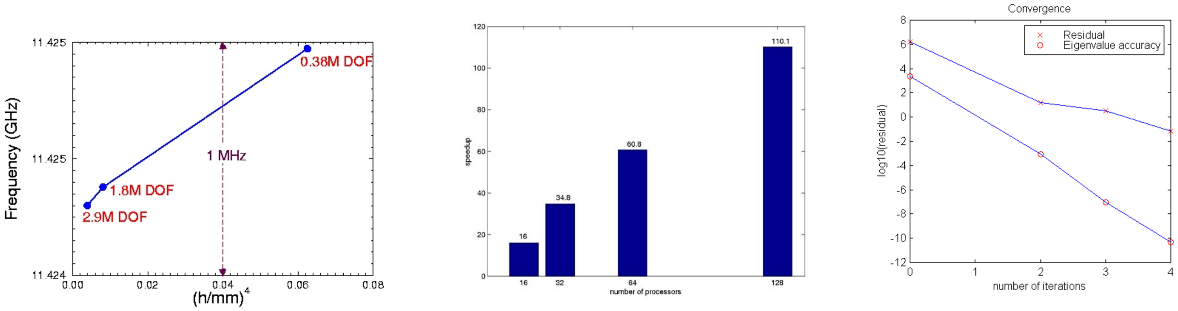

Figure 6a (left) RDDS cavity frequency as function of average mesh size; Fig. 6b (middle) Parallel speedup for 1 million DOF's RDDS model; Fig. 6c (right) Convergence for the same model.

(iv) Adaptive Refinement

The accuracy of solutions by the finite element method depends on how well the finite dimensional space generated by the discretization approximates the solution space. In general, the accuracy improves when the mesh is refined but at additional computational cost. The straightforwrd way is to uniformly subdivide the elements in the mesh and apply this procedure iteratively until the solution converges to a desired degree of accuracy. In this case, the increase in problem size scales with the number of subdivisions for each element. This might work for small problems but for large problems, the method is impractical because the problem size can quickly grow to exceed machine memory or the computation might become too expensive. A more cost effective way to minimize the discretization error is to refine adaptively. By refining only regions of the domain where the error is large, the number of degrees of freedom can be kept to a minimum thereby saving computational cost. Applying this recursively based on local error estimates, an increasingly optimal mesh can be obtained that is well adapted to the solution itself.

Fig. 7 shows three refinements to find the correct peak power density in the waveguide to cavity junction of the PEP-II rf cavity. An accurate calculation is critical for the design of the cooling channels to handle wall heating at high power. These refinements were done manually by remeshing on a single cpu. The problem becomes more difficult when multi-processsors are involved especially on unstructured grids, although the case of a long heterogeneous structure in 2D has been solved. In the absence of the damping manifold, the RDDS structure becomes cylindrically symmetric. Fig. 8 shows the 206-cell detuned structure with cell-to-cell variation through two refinement steps. The approach was to assign each cell to a processor so that communication between processors is limited to the interface elements at the irises. Using an error estimate based on the gradient of the stored energy, an automatic adaptive refinement procedure was able to converge to an accurate solution for a large number of modes (order of thousands) in this structure. The real challenge is parallel automatic mesh refinement on the full 3D 206-cell RDDS structure where domain decomposition is required at each refinement step and this remains an open research problem. Fig, 9 shows two refinement steps of a 12-cell stack on 10 processors repartitioned with ParMetis.

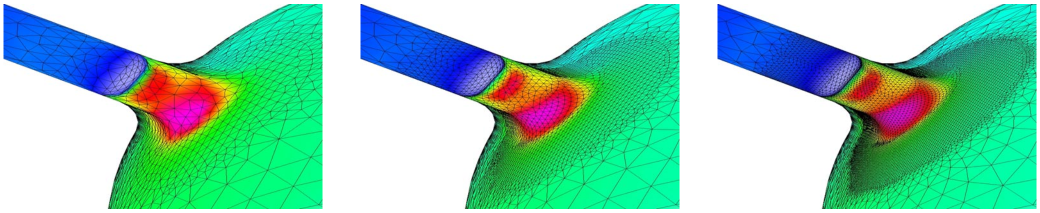

Figure 7. Three refinement steps on the PEP-II rf cavity to find the correct power density on single CPU.

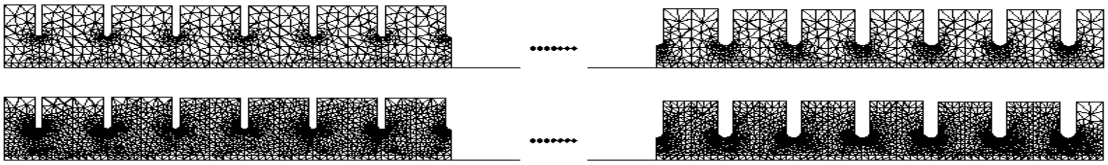

Figure 8. Two automatic adaptive refinement steps on 206-cell detuned structure on 206 processors.

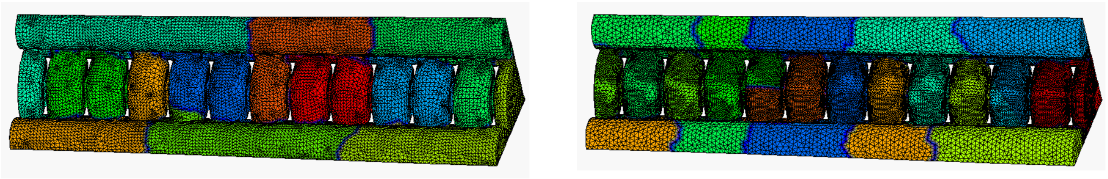

Figure 9 Two refinement steps of a 12-cell RDDS stack on 10 processors repartitioned with ParMetis.

(v) Visualization

Large-scale simulations produce large data sets that require more efficient visualization techniques to analyze because postprocessing with existing graphical packages can present a potential bottleneck. There is ongoing research to explore the use of advanced illumination and interactive methods for the simultaneous display of particles and fields [7]. Fig. 10 shows some preliminary results from this work. The goal is to develop parallel visualization tools capable of handling and interacting with large, time-varying datasets from unstructured meshes that contain complex field vectors and particle trajectories.

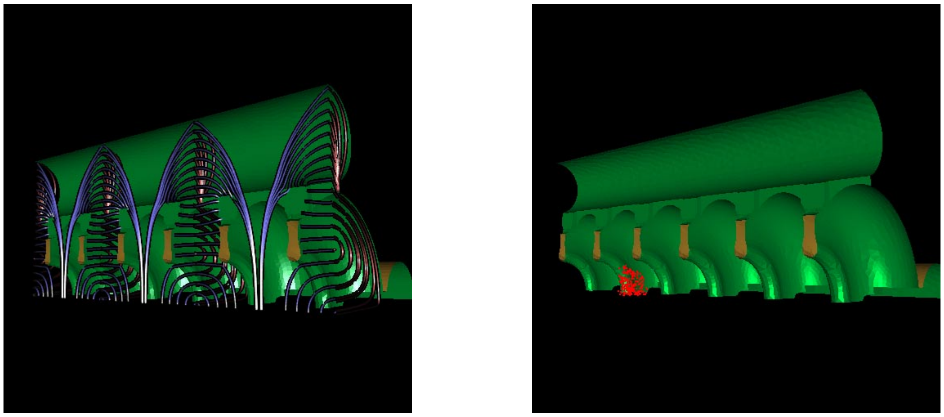

Figure 10a (left) Electric and magnetic field lines in RDDS section; Fig. 10b Particle bunch in same structure at one instant in time.

SUMMARY

Given the great value of particle accelerators to scientific research, and to the nation's health and security, it is imperative that the most advanced high performance computing tools and resources be brought to bear on the challenging and important problems facing the field. Great strides have been made under the Accelerator Grand Challenge towards developing a new generation of simulation tools for modeling accelerators of increasing complexity. Using some of these tools, accelerator designers are solving many problems previously deemed not possible. The effort will continue on a larger scale with the increased funding support from the SciDAC initiative. This paper is an attempt to present a comprehensive overview of the high performance computing program in accelerator physics, and to share some research highlights in the computational electromagnetics area with colleagues in the ACES community.

REFERENCES

1. Zeroth-Order Design Report for the Next Linear Collider, SLAC Report 474, 1996.
2. ParMETIS, Version 2.0, University of Minnesota, 1998, URL: http://www-users.cs.umn.edu/~karypis/metis/parmetis.
3. See these proceedings.
4. SolidEdge, Solid Edge CAD System, Electronic Data Systems, URL: http://www.solid-edge.com.
5. CUBIT, Version 5.0, Sandia National Laboratories, 2001, URL: http://endo.sandia.gov/cubit.
6. PEP-II, An Asymmetric B Factory, Conceptual Design Report, SLAC-418, 1993.
7. G. Schussman et al. Visualizing DIII-D Tokamak magnetic field lines. Proceedings of IEEE Visualization 2000, Salt Lake City, October 2000, pp.501-504.